



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Adding vs. Averaging in Distributed Primal-Dual Optimization

Citation for published version:

Ma, C, Smith, V, Jaggi, M, Jordan, MI, Richtárik, P & Taká, M 2015, 'Adding vs. Averaging in Distributed Primal-Dual Optimization', Paper presented at 32nd International Conference on Machine Learning, Lille, France, 6/07/15 - 11/07/15. <<http://proceedings.mlr.press/v37/mab15.html>>

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Early version, also known as pre-print

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Adding vs. Averaging in Distributed Primal-Dual Optimization

Chenxin Ma*

Industrial and Systems Engineering, Lehigh University, USA

CHM514@LEHIGH.EDU

Virginia Smith*

University of California, Berkeley, USA

VSMITH@BERKELEY.EDU

Martin Jaggi

ETH Zürich, Switzerland

JAGGI@INF.ETHZ.CH

Michael I. Jordan

University of California, Berkeley, USA

JORDAN@CS.BERKELEY.EDU

Peter Richtárik

School of Mathematics, University of Edinburgh, UK

PETER.RICHTARIK@ED.AC.UK

Martin Takáč

Industrial and Systems Engineering, Lehigh University, USA

TAKAC.MT@GMAIL.COM

*Authors contributed equally.

Abstract

Distributed optimization methods for large-scale machine learning suffer from a communication bottleneck. It is difficult to reduce this bottleneck while still efficiently and accurately aggregating partial work from different machines. In this paper, we present a novel generalization of the recent communication-efficient primal-dual framework (CoCoA) for distributed optimization. Our framework, CoCoA⁺, allows for *additive* combination of local updates to the global parameters at each iteration, whereas previous schemes with convergence guarantees only allow conservative averaging. We give stronger (primal-dual) convergence rate guarantees for both CoCoA as well as our new variants, and generalize the theory for both methods to cover non-smooth convex loss functions. We provide an extensive experimental comparison that shows the markedly improved performance of CoCoA⁺ on several real-world distributed datasets, especially when scaling up the number of machines.

1. Introduction

With the wide availability of large datasets that exceed the storage capacity of single machines, distributed optimization methods for machine learning have become increasingly important. Existing methods require significant communication between workers, frequently equaling the amount of local computation (or reading of local data). As a result, distributed machine learning suffers significantly from a communication bottleneck on real world systems, where communication is typically several orders of magnitudes slower than reading data from main memory.

In this work we focus on optimization problems with empirical loss minimization structure, i.e., objectives that are a sum of the loss functions of each datapoint. This includes the most commonly used regularized variants of linear regression and classification methods. For this class of problems, the recently proposed CoCoA approach (Yang, 2013; Jaggi et al., 2014) develops a communication-efficient primal-dual scheme that targets the communication bottleneck, allowing more computation on data-local subproblems native to each machine before communication. By appropriately choosing the amount of local computation per round, this framework allows one to control the trade-off between *communication* and *local computation* based on the systems hardware at hand.

However, the performance of CoCoA (as well as related primal SGD-based methods) is significantly reduced by the

need to average updates between all machines. As the number of machines K grows, the updates get diluted and slowed by $1/K$, e.g., in the case where all machines except one would have already reached the solutions of their respective partial optimization tasks. On the other hand, if the updates are instead added, the algorithms can diverge, as we will observe in the practical experiments below.

To address both described issues, in this paper we develop a novel generalization of the local CoCoA subproblems assigned to each worker, making the framework more powerful in the following sense: Without extra computational cost, the set of locally computed updates from the modified subproblems (one from each machine) can be combined more efficiently between machines. The proposed CoCoA⁺ updates can be aggressively *added* (hence the ‘+’-suffix), which yields much faster convergence both in practice and in theory. This difference is particularly significant as the number of machines K becomes large.

1.1. Contributions

Strong Scaling. To our knowledge, our framework is the first to exhibit favorable *strong scaling* for the class of problems considered, as the number of machines K increases and the data size is kept fixed. More precisely, while the convergence rate of CoCoA degrades as K is increased, the stronger theoretical convergence rate here is – in the worst case – *independent* of K . Our experiments in Section 7 confirm the improved speed of convergence. Since the number of communicated vectors is only one per round and worker, this favorable scaling might be surprising. Indeed, for existing methods, splitting data among more machines generally increases communication requirements (Shamir & Srebro, 2014), which can severely affect overall runtime.

Theoretical Analysis of Non-Smooth Losses. While the existing analysis for CoCoA in (Jaggi et al., 2014) only covered smooth loss functions, here we extend the class of functions where the rates apply, additionally covering, e.g., Support Vector Machines and non-smooth regression variants. We provide a primal-dual convergence rate for both CoCoA as well as our new method CoCoA⁺ in the case of general convex (L -Lipschitz) losses.

Primal-Dual Convergence Rate. Furthermore, we additionally strengthen the rates by showing stronger primal-dual convergence for both algorithmic frameworks, which are almost tight to their objective-only counterparts. Primal-dual rates for CoCoA had not previously been analyzed in the general convex case. Our primal-dual rates allow efficient and practical certificates for the optimization quality, e.g., for stopping criteria. The new rates apply to both smooth and non-smooth losses, and for both CoCoA as well as the extended CoCoA⁺.

Arbitrary Local Solvers. CoCoA as well as CoCoA⁺ allow the use of arbitrary local solvers on each machine.

Experimental Results. We provide a thorough experimental comparison with competing algorithms using several real-world distributed datasets. Our practical results confirm the strong scaling of CoCoA⁺ as the number of machines K grows, while competing methods, including the original CoCoA, slow down significantly with larger K . We implement all algorithms in Spark, and our code is publicly available at: github.com/gingsmith/cocoa.

1.2. History and Related Work

While optimal algorithms for the serial (single machine) case are already well researched and understood, the literature in the distributed setting is relatively sparse. In particular, details on optimal trade-offs between computation and communication, as well as optimization or statistical accuracy, are still widely unclear. For an overview over this currently active research field, we refer the reader to (Balcan et al., 2012; Richtárik & Takáč, 2013; Duchi et al., 2013; Yang, 2013; Liu & Wright, 2014; Fercoq et al., 2014; Jaggi et al., 2014; Shamir & Srebro, 2014; Shamir et al., 2014; Zhang & Lin, 2015; Qu & Richtárik, 2014) and the references therein. We provide a detailed comparison of our proposed framework to the related work in Section 6.

2. Setup

We consider regularized empirical loss minimization problems of the following well-established form:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ \mathcal{P}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n \ell_i(\mathbf{x}_i^T \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right\} \quad (1)$$

Here the vectors $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ represent the training data examples, and the $\ell_i(\cdot)$ are arbitrary convex real-valued loss functions (e.g., hinge loss), possibly depending on label information for the i -th datapoints. The constant $\lambda > 0$ is the regularization parameter.

The above class includes many standard problems of wide interest in machine learning, statistics, and signal processing, including support vector machines, regularized linear and logistic regression, ordinal regression, and others.

Dual Problem, and Primal-Dual Certificates. The conjugate dual of (1) takes following form:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\{ \mathcal{D}(\boldsymbol{\alpha}) := -\frac{1}{n} \sum_{j=1}^n \ell_j^*(-\alpha_j) - \frac{\lambda}{2} \left\| \frac{A\boldsymbol{\alpha}}{\lambda n} \right\|^2 \right\} \quad (2)$$

Here the data matrix $A = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ collects all data-points as its columns, and ℓ_j^* is the conjugate function to ℓ_j . See, e.g., (Shalev-Shwartz & Zhang, 2013c) for several concrete applications.

It is possible to assign for any dual vector $\alpha \in \mathbb{R}^n$ a corresponding primal feasible point

$$\mathbf{w}(\alpha) = \frac{1}{\lambda n} A \alpha \quad (3)$$

The duality gap function is then given by:

$$G(\alpha) := \mathcal{P}(\mathbf{w}(\alpha)) - \mathcal{D}(\alpha) \quad (4)$$

By weak duality, every value $\mathcal{D}(\alpha)$ at a dual candidate α provides a lower bound on every primal value $\mathcal{P}(\mathbf{w})$. The duality gap is therefore a certificate on the approximation quality: The distance to the unknown true optimum $\mathcal{P}(\mathbf{w}^*)$ must always lie within the duality gap, i.e., $G(\alpha) = \mathcal{P}(\mathbf{w}) - \mathcal{D}(\alpha) \geq \mathcal{P}(\mathbf{w}) - \mathcal{P}(\mathbf{w}^*) \geq 0$.

In large-scale machine learning settings like those considered here, the availability of such a computable measure of approximation quality is a significant benefit during training time. Practitioners using classical primal-only methods such as SGD have no means by which to accurately detect if a model has been well trained, as $\mathcal{P}(\mathbf{w}^*)$ is unknown.

Classes of Loss-Functions. To simplify presentation, we assume that all loss functions ℓ_i are non-negative, and

$$\ell_i(0) \leq 1 \quad \forall i \quad (5)$$

Definition 1 (L -Lipschitz continuous loss). A function $\ell_i : \mathbb{R} \rightarrow \mathbb{R}$ is L -Lipschitz continuous if $\forall a, b \in \mathbb{R}$, we have

$$|\ell_i(a) - \ell_i(b)| \leq L|a - b| \quad (6)$$

Definition 2 ($(1/\mu)$ -smooth loss). A function $\ell_i : \mathbb{R} \rightarrow \mathbb{R}$ is $(1/\mu)$ -smooth if it is differentiable and its derivative is $(1/\mu)$ -Lipschitz continuous, i.e., $\forall a, b \in \mathbb{R}$, we have

$$|\ell'_i(a) - \ell'_i(b)| \leq \frac{1}{\mu}|a - b| \quad (7)$$

3. The CoCoA⁺ Algorithm Framework

In this section we present our novel CoCoA⁺ framework. CoCoA⁺ inherits the many benefits of CoCoA as it remains a highly flexible and scalable, communication-efficient framework for distributed optimization. CoCoA⁺ differs algorithmically in that we modify the form of the local subproblems (9) to allow for more aggressive additive updates (as controlled by γ). We will see that these changes allow for stronger convergence guarantees as well as improved empirical performance. Proofs of all statements in this section are given in the supplementary material.

Data Partitioning. We write $\{\mathcal{P}_k\}_{k=1}^K$ for the given partition of the datapoints $[n] := \{1, 2, \dots, n\}$ over the K worker machines. We denote the size of each part by $n_k = |\mathcal{P}_k|$. For any $k \in [K]$ and $\alpha \in \mathbb{R}^n$ we use the notation $\alpha_{[k]} \in \mathbb{R}^{n_k}$ for the vector

$$(\alpha_{[k]})_i := \begin{cases} 0, & \text{if } i \notin \mathcal{P}_k, \\ \alpha_i, & \text{otherwise.} \end{cases}$$

Local Subproblems in CoCoA⁺. We can define a data-local subproblem of the original dual optimization problem (2), which can be solved on machine k and only requires accessing data which is already available locally, i.e., datapoints with $i \in \mathcal{P}_k$. More formally, each machine k is assigned the following local subproblem, depending only on the previous shared primal vector $\mathbf{w} \in \mathbb{R}^d$, and the change in the local dual variables α_i with $i \in \mathcal{P}_k$:

$$\max_{\Delta \alpha_{[k]} \in \mathbb{R}^{n_k}} \mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}; \mathbf{w}, \alpha_{[k]}) \quad (8)$$

where

$$\begin{aligned} \mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}; \mathbf{w}, \alpha_{[k]}) := & -\frac{1}{n} \sum_{i \in \mathcal{P}_k} \ell_i^*(-\alpha_i - (\Delta \alpha_{[k]})_i) \\ & - \frac{1}{K} \frac{\lambda}{2} \|\mathbf{w}\|^2 - \frac{1}{n} \mathbf{w}^T A \Delta \alpha_{[k]} - \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A \Delta \alpha_{[k]} \right\|^2 \end{aligned} \quad (9)$$

Interpretation. The above definition of the local objective functions $\mathcal{G}_k^{\sigma'}$ are such that they closely approximate the global dual objective \mathcal{D} , as we vary the ‘local’ variable $\Delta \alpha_{[k]}$, in the following precise sense:

Lemma 3. For any dual $\alpha, \Delta \alpha \in \mathbb{R}^n$, primal $\mathbf{w} = \mathbf{w}(\alpha)$ and real values γ, σ' satisfying (11), it holds that

$$\begin{aligned} \mathcal{D}\left(\alpha + \gamma \sum_{k=1}^K \Delta \alpha_{[k]}\right) \geq & (1 - \gamma) \mathcal{D}(\alpha) \\ & + \gamma \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}; \mathbf{w}, \alpha_{[k]}) \end{aligned} \quad (10)$$

The role of the parameter σ' is to measure the difficulty of the given data partition. For our purposes, we will see that it must be chosen not smaller than

$$\sigma' \geq \sigma'_{\min} := \gamma \max_{\alpha \in \mathbb{R}^n} \frac{\|A \alpha\|^2}{\sum_{k=1}^K \|A \alpha_{[k]}\|^2} \quad (11)$$

In the following lemma, we show that this parameter can be upper-bounded by γK , which is trivial to calculate for all values $\gamma \in \mathbb{R}$. We show experimentally (Section 7) that this safe upper bound for σ' has a minimal effect on the overall performance of the algorithm. Our main theorems below show convergence rates dependent on $\gamma \in [\frac{1}{K}, 1]$, which we refer to as the *aggregation parameter*.

Lemma 4. The choice of $\sigma' := \gamma K$ is valid for (11), i.e.,

$$\gamma K \geq \sigma'_{\min}$$

Notion of Approximation Quality of the Local Solver.

Assumption 1 (Θ -approximate solution). We assume that there exists $\Theta \in [0, 1]$ such that $\forall k \in [K]$, the local solver at any outer iteration t produces a (possibly) randomized approximate solution $\Delta \alpha_{[k]}$, which satisfies

$$\begin{aligned} \mathbb{E}[\mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}; \mathbf{w}, \alpha_{[k]})] & \quad (12) \\ \leq \Theta \left(\mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(0; \mathbf{w}, \alpha_{[k]}) \right), \end{aligned}$$

where

$$\Delta \alpha_{[k]}^* \in \arg \max_{\Delta \alpha \in \mathbb{R}^n} \mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}; \mathbf{w}, \alpha_{[k]}) \quad \forall k \in [K] \quad (13)$$

We are now ready to describe the CoCoA⁺ framework, shown in Algorithm 1. The crucial difference compared to the existing CoCoA algorithm (Jaggi et al., 2014) is the more general local subproblem, as defined in (9), as well as the aggregation parameter γ . These modifications allow the option of directly adding updates to the global vector \mathbf{w} .

Algorithm 1 CoCoA⁺ Framework

- 1: **Input:** Datapoints A distributed according to partition $\{\mathcal{P}_k\}_{k=1}^K$. Aggregation parameter $\gamma \in (0, 1]$, subproblem parameter σ' for the local subproblems $\mathcal{G}_k^{\sigma'}(\Delta \alpha_{[k]}; \mathbf{w}, \alpha_{[k]})$ for each $k \in [K]$. Starting point $\alpha^{(0)} := \mathbf{0} \in \mathbb{R}^n$, $\mathbf{w}^{(0)} := \mathbf{0} \in \mathbb{R}^d$.
 - 2: **for** $t = 0, 1, 2, \dots$ **do**
 - 3: **for** $k \in \{1, 2, \dots, K\}$ **in parallel over computers do**
 - 4: call the local solver, computing a Θ -approximate solution $\Delta \alpha_{[k]}$ of the local subproblem (9)
 - 5: update $\alpha_{[k]}^{(t+1)} := \alpha_{[k]}^{(t)} + \gamma \Delta \alpha_{[k]}$
 - 6: return $\Delta \mathbf{w}_k := \frac{1}{\lambda n} A \Delta \alpha_{[k]}$
 - 7: **end for**
 - 8: reduce $\mathbf{w}^{(t+1)} := \mathbf{w}^{(t)} + \gamma \sum_{k=1}^K \Delta \mathbf{w}_k$. (14)
 - 9: **end for**
-

4. Convergence Guarantees

Before being able to state our main convergence results, we introduce some useful quantities and the following main lemma characterizing the effect of iterations of Algorithm 1, for any chosen internal local solver.

Lemma 5. Let ℓ_i^* be strongly¹ convex with convexity parameter $\mu \geq 0$ with respect to the norm $\|\cdot\|$, $\forall i \in [n]$. Then for all iterations t of Algorithm 1 under Assumption 1, and any $s \in [0, 1]$, it holds that

$$\mathbb{E}[\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^{(t)})] \geq \gamma(1 - \Theta) \left(sG(\alpha^{(t)}) - \frac{\sigma'}{2\lambda} \left(\frac{s}{n} \right)^2 R^{(t)} \right), \quad (15)$$

where

$$R^{(t)} := -\frac{\lambda \mu n(1-s)}{\sigma' s} \|\mathbf{u}^{(t)} - \alpha^{(t)}\|^2 + \sum_{k=1}^K \|A(\mathbf{u}^{(t)} - \alpha^{(t)})_{[k]}\|^2, \quad (16)$$

for $\mathbf{u}^{(t)} \in \mathbb{R}^n$ with

$$-u_i^{(t)} \in \partial \ell_i(\mathbf{w}(\alpha^{(t)})^T \mathbf{x}_i). \quad (17)$$

¹Note that the case of weakly convex $\ell_i^*(\cdot)$ is explicitly allowed here as well, as the Lemma holds for the case $\mu = 0$.

The following Lemma provides a uniform bound on $R^{(t)}$:

Lemma 6. If ℓ_i are L -Lipschitz continuous for all $i \in [n]$, then

$$\forall t : R^{(t)} \leq 4L^2 \underbrace{\sum_{k=1}^K \sigma_k n_k}_{=: \sigma}, \quad (18)$$

where

$$\sigma_k := \max_{\alpha_{[k]} \in \mathbb{R}^n} \frac{\|A \alpha_{[k]}\|^2}{\|\alpha_{[k]}\|^2}. \quad (19)$$

Remark 7. If all data-points \mathbf{x}_i are normalized such that $\|\mathbf{x}_i\| \leq 1 \forall i \in [n]$, then $\sigma_k \leq |\mathcal{P}_k| = n_k$. Furthermore, if we assume that the data partition is balanced, i.e., that $n_k = n/K$ for all k , then $\sigma \leq n^2/K$. This can be used to bound the constants $R^{(t)}$, above, as $R^{(t)} \leq \frac{4L^2 n^2}{K}$.

4.1. Primal-Dual Convergence for General Convex Losses

The following theorem shows the convergence for non-smooth loss functions, in terms of objective values as well as primal-dual gap. The analysis in (Jaggi et al., 2014) only covered the case of smooth loss functions.

Theorem 8. Consider Algorithm 1 with Assumption 1. Let $\ell_i(\cdot)$ be L -Lipschitz continuous, and $\epsilon_G > 0$ be the desired duality gap (and hence an upper-bound on primal sub-optimality). Then after T iterations, where

$$T \geq T_0 + \max\left\{\left\lceil \frac{1}{\gamma(1-\Theta)} \right\rceil, \frac{4L^2 \sigma \sigma'}{\lambda n^2 \epsilon_G \gamma(1-\Theta)}\right\}, \quad (20)$$

$$T_0 \geq t_0 + \left(\frac{2}{\gamma(1-\Theta)} \left(\frac{8L^2 \sigma \sigma'}{\lambda n^2 \epsilon_G} - 1 \right) \right)_+,$$

$$t_0 \geq \max(0, \left\lceil \frac{1}{\gamma(1-\Theta)} \log\left(\frac{2\lambda n^2 (\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(0)}))}{4L^2 \sigma \sigma'}\right) \right\rceil),$$

we have that the expected duality gap satisfies

$$\mathbb{E}[\mathcal{P}(\mathbf{w}(\bar{\alpha})) - \mathcal{D}(\bar{\alpha})] \leq \epsilon_G,$$

at the averaged iterate

$$\bar{\alpha} := \frac{1}{T-T_0} \sum_{t=T_0+1}^{T-1} \alpha^{(t)}. \quad (21)$$

The following corollary of the above theorem clarifies our main result: The more aggressive adding of the partial updates, as compared averaging, offers a very significant improvement in terms of total iterations needed. While the convergence in the ‘adding’ case becomes independent of the number of machines K , the ‘averaging’ regime shows the known degradation of the rate with growing K , which is a major drawback of the original CoCoA algorithm. This important difference in the convergence speed is not a theoretical artifact but also confirmed in our practical experiments below for different K , as shown e.g. in Figure 2.

We further demonstrate below that by choosing γ and σ' accordingly, we can still recover the original CoCoA algorithm and its rate.

Corollary 9. Assume that all datapoints \mathbf{x}_i are bounded as $\|\mathbf{x}_i\| \leq 1$ and that the data partition is balanced, i.e. that $n_k = n/K$ for all k . We consider two different possible choices of the aggregation parameter γ :

- (CoCoA Averaging, $\gamma := \frac{1}{K}$): In this case, $\sigma' := 1$ is a valid choice which satisfies (11). Then using $\sigma \leq n^2/K$ in light of Remark 7, we have that T iterations are sufficient for primal-dual accuracy ϵ_G , with

$$\begin{aligned} T &\geq T_0 + \max\left\{\left\lceil \frac{K}{1-\Theta} \right\rceil, \frac{4L^2}{\lambda\epsilon_G(1-\Theta)}\right\}, \\ T_0 &\geq t_0 + \left(\frac{2K}{1-\Theta} \left(\frac{8L^2}{\lambda K \epsilon_G} - 1\right)\right)_+, \\ t_0 &\geq \max(0, \left\lceil \frac{K}{1-\Theta} \log\left(\frac{2\lambda(\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(0)}))}{4KL^2}\right) \right\rceil) \end{aligned}$$

Hence the more machines K , the more iterations are needed (in the worst case).

- (CoCoA⁺ Adding, $\gamma := 1$): In this case, the choice of $\sigma' := K$ satisfies (11). Then using $\sigma \leq n^2/K$ in light of Remark 7, we have that T iterations are sufficient for primal-dual accuracy ϵ_G , with

$$\begin{aligned} T &\geq T_0 + \max\left\{\left\lceil \frac{1}{1-\Theta} \right\rceil, \frac{4L^2}{\lambda\epsilon_G(1-\Theta)}\right\}, \\ T_0 &\geq t_0 + \left(\frac{2}{1-\Theta} \left(\frac{8L^2}{\lambda\epsilon_G} - 1\right)\right)_+, \\ t_0 &\geq \max(0, \left\lceil \frac{1}{1-\Theta} \log\left(\frac{2\lambda n(\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(0)}))}{4KL^2}\right) \right\rceil) \end{aligned}$$

This is significantly better than the averaging case.

In practice, we usually have $\sigma \ll n^2/K$, and hence the actual convergence rate can be much better than the proven worst-case bound. Table 1 shows that the actual value of σ is typically between one and two orders of magnitudes smaller compared to our used upper-bound n^2/K .

Table 1. The ratio of upper-bound $\frac{n^2}{K}$ divided by the true value of the parameter σ , for some real datasets.

K	16	32	64	128	256	512
news	15.483	14.933	14.278	13.390	12.074	10.252
real-sim	42.127	36.898	30.780	23.814	16.965	11.835
rcv1	40.138	23.827	28.204	21.792	16.339	11.099
K	256	512	1024	2048	4096	8192
covtype	17.277	17.260	17.239	16.948	17.238	12.729

4.2. Primal-Dual Convergence for Smooth Losses

The following theorem shows the convergence for smooth losses, in terms of the objective as well as primal-dual gap.

Theorem 10. Assume the loss functions ℓ_i are $(1/\mu)$ -smooth $\forall i \in [n]$. We define $\sigma_{\max} = \max_{k \in [K]} \sigma_k$. Then after T iterations of Algorithm 1, with

$$T \geq \frac{1}{\gamma(1-\Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \log \frac{1}{\epsilon_D},$$

it holds that

$$\mathbb{E}[\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(T)})] \leq \epsilon_D.$$

Furthermore, after T iterations with

$$T \geq \frac{1}{\gamma(1-\Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \log \left(\frac{1}{\gamma(1-\Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \frac{1}{\epsilon_G} \right),$$

we have the expected duality gap

$$\mathbb{E}[\mathcal{P}(\mathbf{w}(\alpha^{(T)})) - \mathcal{D}(\alpha^{(T)})] \leq \epsilon_G.$$

The following corollary is analogous to Corollary 9, but for the case of smooth losses. It again shows that while the CoCoA variant degrades with the increase of the number of machines K , the CoCoA⁺ rate is independent of K .

Corollary 11. Assume that all datapoints \mathbf{x}_i are bounded as $\|\mathbf{x}_i\| \leq 1$ and that the data partition is balanced, i.e., that $n_k = n/K$ for all k . We again consider the same two different possible choices of the aggregation parameter γ :

- (CoCoA Averaging, $\gamma := \frac{1}{K}$): In this case, $\sigma' := 1$ is a valid choice which satisfies (11). Then using $\sigma_{\max} \leq n_k = n/K$ in light of Remark 7, we have that T iterations are sufficient for suboptimality ϵ_D , with

$$T \geq \frac{1}{1-\Theta} \frac{\lambda\mu K + 1}{\lambda\mu} \log \frac{1}{\epsilon_D}$$

Hence the more machines K , the more iterations are needed (in the worst case).

- (CoCoA⁺ Adding, $\gamma := 1$): In this case, the choice of $\sigma' := K$ satisfies (11). Then using $\sigma_{\max} \leq n_k = n/K$ in light of Remark 7, we have that T iterations are sufficient for suboptimality ϵ_D , with

$$T \geq \frac{1}{1-\Theta} \frac{\lambda\mu + 1}{\lambda\mu} \log \frac{1}{\epsilon_D}$$

This is significantly better than the averaging case. Both rates hold analogously for the duality gap.

4.3. Comparison with Original CoCoA

Remark 12. If we choose averaging ($\gamma := \frac{1}{K}$) for aggregating the updates, together with $\sigma' := 1$, then the resulting Algorithm 1 is identical to CoCoA analyzed in (Jaggi et al., 2014). However, they only provide convergence for smooth loss functions ℓ_i and have guarantees for dual suboptimality and not the duality gap. Formally, when $\sigma' = 1$, the subproblems (9) will differ from the original dual $\mathcal{D}(\cdot)$ only by an additive constant, which does not affect the local optimization algorithms used within CoCoA.

5. SDCA as an Example Local Solver

We have shown convergence rates for Algorithm 1, depending solely on the approximation quality Θ of the used local

solver (Assumption 1). Any chosen local solver in each round receives the local α variables as an input, as well as a shared vector $\mathbf{w} \stackrel{(3)}{=} \mathbf{w}(\alpha)$ being compatible with the last state of all global $\alpha \in \mathbb{R}^n$ variables.

As an illustrative example for a local solver, Algorithm 2 below summarizes randomized coordinate ascent (SDCA) applied on the local subproblem (9). The following two Theorems (13, 14) characterize the local convergence for both smooth and non-smooth functions. In all the results we will use $r_{\max} := \max_{i \in [n]} \|\mathbf{x}_i\|^2$.

Algorithm 2 LOCALSDCA ($\mathbf{w}, \alpha_{[k]}, k, H$)

```

1: Input:  $\alpha_{[k]}, \mathbf{w} = \mathbf{w}(\alpha)$ 
2: Data: Local  $\{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{P}_k}$ 
3: Initialize:  $\Delta\alpha_{[k]}^{(0)} := \mathbf{0} \in \mathbb{R}^n$ 
4: for  $h = 0, 1, \dots, H - 1$  do
5:   choose  $i \in \mathcal{P}_k$  uniformly at random
6:    $\delta_i^* := \arg \max_{\delta_i \in \mathbb{R}} \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^{(h)} + \delta_i \mathbf{e}_i; \mathbf{w}, \alpha_{[k]})$ 
7:    $\Delta\alpha_{[k]}^{(h+1)} := \Delta\alpha_{[k]}^{(h)} + \delta_i^* \mathbf{e}_i$ 
8: end for
9: Output:  $\Delta\alpha_{[k]}^{(H)}$ 
    
```

Theorem 13. Assume the functions ℓ_i are $(1/\mu)$ -smooth for $i \in [n]$. Then Assumption 1 on the local approximation quality Θ is satisfied for LOCALSDCA as given in Algorithm 2, if we choose the number of inner iterations H as

$$H \geq n_k \frac{\sigma' r_{\max} + \lambda n \mu}{\lambda n \mu} \log \frac{1}{\Theta}. \quad (22)$$

Theorem 14. Assume the functions ℓ_i are L -Lipschitz for $i \in [n]$. Then Assumption 1 on the local approximation quality Θ is satisfied for LOCALSDCA as given in Algorithm 2, if we choose the number of inner iterations H as

$$H \geq n_k \left(\frac{1 - \Theta}{\Theta} + \frac{\sigma' r_{\max}}{2\Theta \lambda n^2} \frac{\|\Delta\alpha_{[k]}^*\|^2}{\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \cdot) - \mathcal{G}_k^{\sigma'}(\mathbf{0}; \cdot)} \right). \quad (23)$$

Remark 15. Between the different regimes allowed in CoCoA⁺ (ranging between averaging and adding the updates) the computational cost for obtaining the required local approximation quality varies with the choice of σ' . From the above worst-case upper bound, we note that the cost can increase with σ' , as aggregation becomes more aggressive. However, as we will see in the practical experiments in Section 7 below, the additional cost is negligible compared to the gain in speed from the different aggregation, when measured on real datasets.

6. Discussion and Related Work

SGD-based Algorithms. For the empirical loss minimization problems of interest here, stochastic subgradient

descent (SGD) based methods are well-established. Several distributed variants of SGD have been proposed, many of which build on the idea of a parameter server (Niu et al., 2011; Liu et al., 2014; Duchi et al., 2013). The downside of this approach, even when carefully implemented, is that the amount of required communication is equal to the amount of data read locally (e.g., mini-batch SGD with a batch size of 1 per worker). These variants are in practice not competitive with the more communication-efficient methods considered here, which allow more local updates per round.

One-Shot Communication Schemes. At the other extreme, there are distributed methods using only a single round of communication, such as (Zhang et al., 2013; Zinkevich et al., 2010; Mann et al., 2009; McWilliams et al., 2014). These require additional assumptions on the partitioning of the data, and furthermore can not guarantee convergence to the optimum solution for all regularizers, as shown in, e.g., (Shamir et al., 2014). (Balcan et al., 2012) shows additional relevant lower bounds on the minimum number of communication rounds necessary for a given approximation quality for similar machine learning problems.

Mini-Batch Methods. Mini-batch methods are more flexible and lie within these two communication vs. computation extremes. However, mini-batch versions of both SGD and coordinate descent (CD) (Richtárik & Takáč, 2013; Shalev-Shwartz & Zhang, 2013b; Yang, 2013; Qu & Richtárik, 2014; Qu et al., 2014) suffer from their convergence rate degrading towards the rate of batch gradient descent as the size of the mini-batch is increased. This follows because mini-batch updates are made based on the outdated previous parameter vector \mathbf{w} , in contrast to methods that allow immediate local updates like CoCoA. Furthermore, the aggregation parameter for mini-batch methods is harder to tune, as it can lie anywhere in the order of mini-batch size. In the CoCoA setting, the parameter lies in the smaller range given by K . Our CoCoA⁺ extension avoids needing to tune this parameter entirely, by adding.

Methods Allowing Local Optimization. Developing methods that allow for local optimization requires carefully devising data-local subproblems to be solved after each communication round. (Shamir et al., 2014; Zhang & Lin, 2015) have proposed distributed Newton-type algorithms in this spirit. However, the subproblems must be solved to high accuracy for convergence to hold, which is often prohibitive as the size of the data on one machine is still relatively large. In contrast, the CoCoA framework (Jaggi et al., 2014) allows using any local solver of weak local approximation quality in each round. By making use of the primal-dual structure in the line of work of (Yu et al., 2012; Pechyony et al., 2011; Yang, 2013; Lee & Roth, 2015), the CoCoA and CoCoA⁺ frameworks also allow more control over the aggregation of updates between ma-

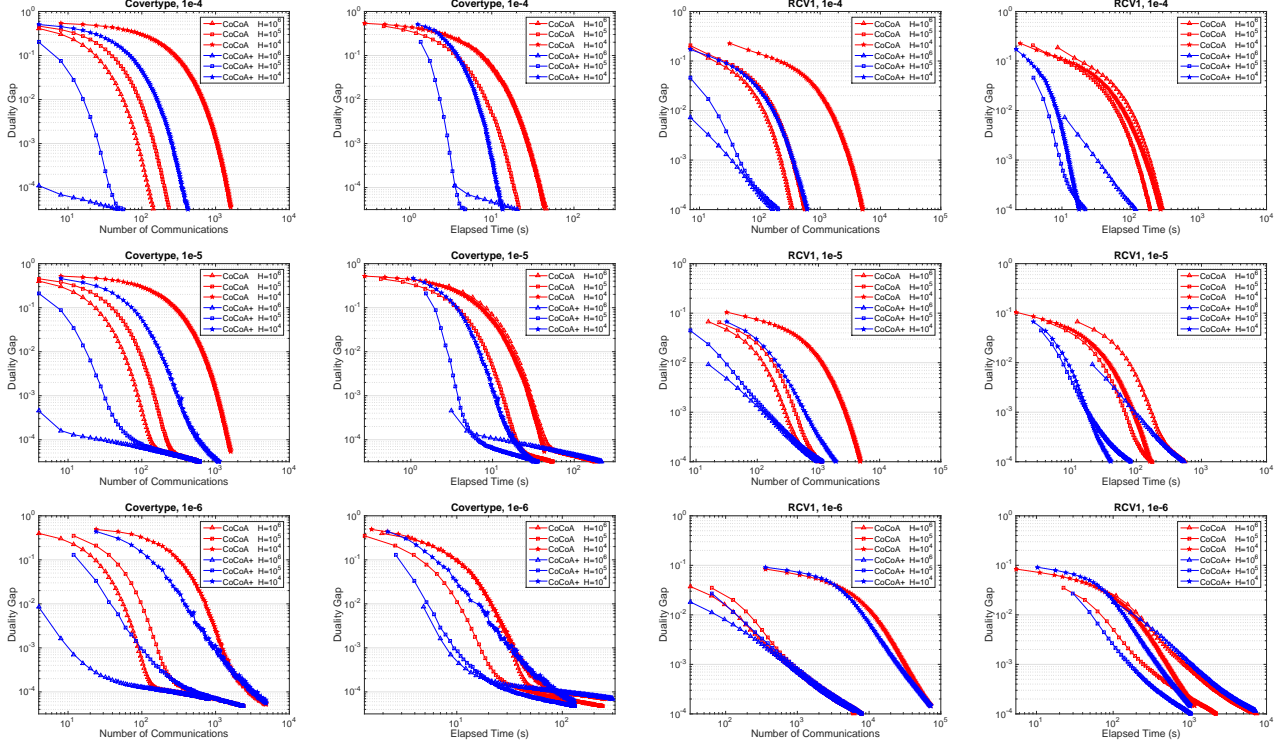


Figure 1. Duality gap vs. the number of communicated vectors, as well as duality gap vs. elapsed time in seconds for two datasets: Covertypes (left, $K=4$) and RCV1 (right, $K=8$). Both are shown on a log-log scale, and for three different values of regularization ($\lambda=1e-4; 1e-5; 1e-6$). Each plot contains a comparison of CoCoA (red) and CoCoA⁺ (blue), for three different values of H , the number of local iterations performed per round. For all plots, across all values of λ and H , we see that CoCoA⁺ converges to the optimal solution faster than CoCoA, in terms of both the number of communications and the elapsed time.

chines. The practical variant DisDCA-p proposed in (Yang, 2013) allows additive updates but is restricted to SDCA updates, and was proposed without convergence guarantees. DisDCA-p can be recovered as a special case of the CoCoA⁺ framework when using SDCA as a local solver, if $n_k = n/K$ and $\sigma' := K$, see Appendix C. The theory presented here also therefore covers that method.

ADMM. An alternative approach to distributed optimization is to use the alternating direction method of multipliers (ADMM), as used for distributed SVM training in, e.g., (Forero et al., 2010). This uses a penalty parameter balancing between the equality constraint \mathbf{w} and the optimization objective (Boyd et al., 2011). However, the known convergence rates for ADMM are weaker than the more problem-tailored methods mentioned previously, and the choice of the penalty parameter is often unclear.

Batch Proximal Methods. In spirit, for the special case of adding ($\gamma = 1$), CoCoA⁺ resembles a batch proximal method, using the separable approximation (9) instead of the original dual (2). Known batch proximal methods require high accuracy subproblem solutions, and don’t allow arbitrary solvers of weak accuracy Θ such as we do here.

7. Numerical Experiments

We present experiments on several large real-world distributed datasets. We show that CoCoA⁺ converges faster in terms of total rounds as well as elapsed time as compared to CoCoA in all cases, despite varying: the dataset, values of regularization, batch size, and cluster size (Section 7.2). In Section 7.3 we demonstrate that this performance translates to orders of magnitude improvement in convergence when scaling up the number of machines K , as compared to CoCoA as well as to several other state-of-the-art methods. Finally, in Section 7.4 we investigate the impact of the local subproblem parameter σ' in the CoCoA⁺ framework.

Table 2. Datasets for Numerical Experiments.

Dataset	n	d	Sparsity
covertypes	522,911	54	22.22%
epsilon	400,000	2,000	100%
RCV1	677,399	47,236	0.16%

7.1. Implementation Details

We implement all algorithms in Apache Spark (Zaharia et al., 2012) and run them on m3.large Amazon EC2 instances, applying each method to the binary hinge-loss sup-

port vector machine. The analysis for this non-smooth loss was not covered in (Jaggi et al., 2014) but has been captured here, and thus is both theoretically and practically justified. The used datasets are summarized in Table 2.

For illustration and ease of comparison, we here use SDCA (Shalev-Shwartz & Zhang, 2013c) as the local solver for both CoCoA and CoCoA⁺. Note that in this special case, and if additionally $\sigma' := K$, and if the partitioning $n_k = n/K$ is balanced, once can show that the CoCoA⁺ framework reduces to the practical variant of DisDCA (Yang, 2013) (which had no convergence guarantees so far). We include more details on the connection in Appendix C.

7.2. Comparison of CoCoA⁺ and CoCoA

We compare the CoCoA⁺ and CoCoA frameworks directly using two datasets (Coverttype and RCV1) across various values of λ , the regularizer, in Figure 1. For each value of λ we consider both methods with different values of H , the number of local iterations performed before communicating to the master. For all runs of CoCoA⁺ we use the safe upper bound of γK for σ' . In terms of both the total number of communications made and the elapsed time, CoCoA⁺ (shown in blue) converges to the optimal solution faster than CoCoA (red). The discrepancy is larger for greater values of λ , where the strongly convex regularizer has more of an impact and the problem difficulty is reduced. We also see a greater performance gap for smaller values of H , where there is frequent communication between the machines and the master, and changes between the algorithms therefore play a larger role.

7.3. Scaling the Number of Machines K

In Figure 2 we demonstrate the ability of CoCoA⁺ to scale with an increasing number of machines K . The experiments confirm the ability of strong scaling of the new method, as predicted by our theory in Section 4, in contrast to the competing methods. Unlike CoCoA, which becomes linearly slower when increasing the number of machines, the performance of CoCoA⁺ improves with additional machines, only starting to degrade slightly once $K=16$ for the RCV1 dataset.

7.4. Impact of the Subproblem Parameter σ'

Finally, in Figure 3, we consider the effect of the choice of the subproblem parameter σ' on convergence. We plot both the number of communications and clock time on a log-log scale for the RCV1 dataset with $K=8$ and $H=1e4$. For $\gamma = 1$ (the most aggressive variant of CoCoA⁺ in which updates are added) we consider several different values of σ' , ranging from 1 to 8. The value $\sigma'=8$ represents the safe upper bound of γK . The optimal convergence occurs around $\sigma'=4$, and diverges for $\sigma' \leq 2$. Notably, we

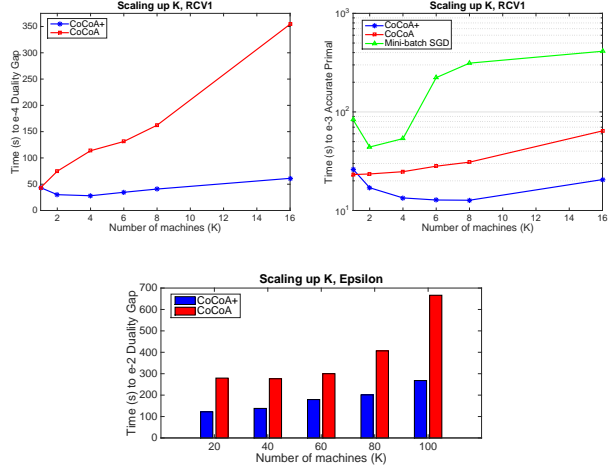


Figure 2. The effect of increasing K on the time (s) to reach an $\epsilon_{\mathcal{D}}$ -accurate solution. We see that CoCoA⁺ converges twice as fast as CoCoA on 100 machines for the Epsilon dataset, and nearly 7 times as quickly for the RCV1 dataset. Mini-batch SGD converges an order of magnitude more slowly than both methods.

see that the easy to calculate upper bound of $\sigma' := \gamma K$ (as given by Lemma 4) has only slightly worse performance than best possible subproblem parameter in our setting.

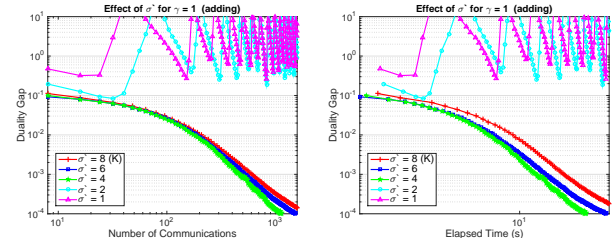


Figure 3. The effect of σ' on convergence of CoCoA⁺ for the RCV1 dataset distributed across $K=8$ machines. Decreasing σ' improves performance in terms of communication and overall run time until a certain point, after which the algorithm diverges. The “safe” upper bound of $\sigma' := K=8$ has only slightly worse performance than the practically best “un-safe” value of σ' .

8. Conclusion

In conclusion, we present a novel framework CoCoA⁺ that allows for fast and communication-efficient *additive aggregation* in distributed algorithms for primal-dual optimization. We analyze the theoretical performance of this method, giving strong primal-dual convergence rates with outer iterations scaling independently of the number of machines. We extended our theory to allow for non-smooth losses. Our experimental results show significant speedups over previous methods, including the original CoCoA framework as well as other state-of-the-art methods.

Acknowledgments. We thank Ching-pei Lee and an anonymous reviewer for several helpful insights and comments.

References

- Balcan, M.-F., Blum, A., Fine, S., and Mansour, Y. Distributed Learning, Communication Complexity and Privacy. In *COLT*, pp. 26.1–26.22, 2012.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- Duchi, J. C., Jordan, M. I., and McMahan, H. B. Estimation, Optimization, and Parallelism when Data is Sparse. In *NIPS*, 2013.
- Fercoq, O. and Richtárik, P. Accelerated, parallel and proximal coordinate descent. *arXiv:1312.5799*, 2013.
- Fercoq, O., Qu, Z., Richtárik, P., and Takáč, M. Fast distributed coordinate descent for non-strongly convex losses. *IEEE Workshop on Machine Learning for Signal Processing*, 2014.
- Forero, P. A., Cano, A., and Giannakis, G. B. Consensus-Based Distributed Support Vector Machines. *JMLR*, 11: 1663–1707, 2010.
- Jaggi, M., Smith, V., Takáč, M., Terhorst, J., Krishnan, S., Hofmann, T., and Jordan, M. I. Communication-efficient distributed dual coordinate ascent. In *NIPS*, 2014.
- Lee, C.-P. and Roth, D. Distributed Box-Constrained Quadratic Optimization for Dual Linear SVM. In *ICML*, 2015.
- Liu, J. and Wright, S. J. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *arXiv:1403.3862*, 2014.
- Liu, J., Wright, S. J., Ré, C., Bittorf, V., and Sridhar, S. An Asynchronous Parallel Stochastic Coordinate Descent Algorithm. In *ICML*, 2014.
- Lu, Z. and Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *arXiv preprint arXiv:1305.4723*, 2013.
- Mann, G., McDonald, R., Mohri, M., Silberman, N., and Walker, D. D. Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models. *NIPS*, 2009.
- Mareček, J., Richtárik, P., and Takáč, M. Distributed block coordinate descent for minimizing partially separable functions. *arXiv:1406.0238*, 2014.
- McWilliams, B., Heinze, C., Meinshausen, N., Krummenacher, G., and Vanchinathan, H. P. LOCO: Distributing Ridge Regression with Random Projections. *arXiv stat.ML*, June 2014.
- Niu, F., Recht, B., Ré, C., and Wright, S. J. Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In *NIPS*, 2011.
- Pechyony, D., Shen, L., and Jones, R. Solving Large Scale Linear SVM with Distributed Block Minimization. In *NIPS Workshop on Big Learning*, 2011.
- Qu, Z. and Richtárik, P. Coordinate descent with arbitrary sampling I: Algorithms and complexity. *arXiv:1412.8060*, 2014.
- Qu, Z., Richtárik, P., and Zhang, T. Randomized dual coordinate ascent with arbitrary sampling. *arXiv:1411.5873*, 2014.
- Richtárik, P. and Takáč, M. Distributed coordinate descent method for learning with big data. *arXiv preprint arXiv:1310.2059*, 2013.
- Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, April 2014.
- Richtárik, P. and Takáč, M. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, pp. 1–52, 2015.
- Shalev-Shwartz, S. and Zhang, T. Accelerated mini-batch stochastic dual coordinate ascent. In *NIPS*, 2013a.
- Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *arXiv:1309.2375*, 2013b.
- Shalev-Shwartz, S. and Zhang, T. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *JMLR*, 14:567–599, 2013c.
- Shamir, O. and Srebro, N. Distributed Stochastic Optimization and Learning . In *Allerton*, 2014.
- Shamir, O., Srebro, N., and Zhang, T. Communication efficient distributed optimization using an approximate newton-type method. In *ICML*, 2014.
- Tappenden, R., Takáč, M., and Richtárik, P. On the complexity of parallel coordinate descent. Technical report, 2015. ERGO 15-001, University of Edinburgh.
- Yang, T. Trading Computation for Communication: Distributed Stochastic Dual Coordinate Ascent. In *NIPS*, 2013.
- Yang, T., Zhu, S., Jin, R., and Lin, Y. On Theoretical Analysis of Distributed Stochastic Dual Coordinate Ascent. *arXiv:1312.1031*, 2013.

- Yu, H.-F., Hsieh, C.-J., Chang, K.-W., and Lin, C.-J. Large Linear Classification When Data Cannot Fit in Memory. *TKDD*, 5(4):1–23, 2012.
- Zaharia, M., Chowdhury, M., Das, T., Dave, A., McCauley, M., Franklin, M. J., Shenker, S., and Stoica, I. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. In *NSDI*, 2012.
- Zhang, Y. and Lin, X. DiSCO: Distributed Optimization for Self-Concordant Empirical Loss. In *ICML*, pp. 362–370, 2015.
- Zhang, Y., Duchi, J. C., and Wainwright, M. J. Communication-Efficient Algorithms for Statistical Optimization. *JMLR*, 14:3321–3363, 2013.
- Zinkevich, M. A., Weimer, M., Smola, A. J., and Li, L. Parallelized Stochastic Gradient Descent. *NIPS*, 2010.

Appendix

A. Technical Lemmas

Lemma 16 (Lemma 21 in (Shalev-Shwartz & Zhang, 2013c)). *Let $\ell_i : \mathbb{R} \rightarrow \mathbb{R}$ be an L -Lipschitz continuous. Then for any real value a with $|a| > L$ we have that $\ell_i^*(a) = \infty$.*

Lemma 17. *Assuming the loss functions ℓ_i are bounded by $\ell_i(0) \leq 1$ for all $i \in [n]$ (as we have assumed in (5) above), then for the zero vector $\alpha^{(0)} := \mathbf{0} \in \mathbb{R}^n$, we have*

$$\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(0)}) = \mathcal{D}(\alpha^*) - \mathcal{D}(\mathbf{0}) \leq 1. \quad (24)$$

Proof. For $\alpha := \mathbf{0} \in \mathbb{R}^n$, we have $\mathbf{w}(\alpha) = \frac{1}{\lambda n} A \alpha = \mathbf{0} \in \mathbb{R}^d$. Therefore, by definition of the dual objective \mathcal{D} given in (2),

$$0 \leq \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha) \leq \mathcal{P}(\mathbf{w}(\alpha)) - \mathcal{D}(\alpha) = 0 - \mathcal{D}(\alpha) \stackrel{(5),(2)}{\leq} 1. \quad \square$$

B. Proofs

B.1. Proof of Lemma 3

Indeed, we have

$$\mathcal{D}(\alpha + \gamma \sum_{k=1}^K \Delta \alpha_{[k]}) = \underbrace{-\frac{1}{n} \sum_{i=1}^n \ell_i^*(-\alpha_i - \gamma (\sum_{k=1}^K \Delta \alpha_{[k]})_i)}_A - \underbrace{\frac{\lambda}{2} \left\| \frac{1}{\lambda n} A (\alpha + \gamma \sum_{k=1}^K \Delta \alpha_{[k]}) \right\|^2}_B. \quad (25)$$

Now, let us bound the terms A and B separately. We have

$$\begin{aligned} A &= -\frac{1}{n} \sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} \ell_i^*(-\alpha_i - \gamma (\Delta \alpha_{[k]})_i) \right) = -\frac{1}{n} \sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} \ell_i^*(-(1-\gamma)\alpha_i - \gamma(\alpha + \Delta \alpha_{[k]})_i) \right) \\ &\geq -\frac{1}{n} \sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} (1-\gamma) \ell_i^*(-\alpha_i) + \gamma \ell_i^*(-(\alpha + \Delta \alpha_{[k]})_i) \right). \end{aligned}$$

Where the last inequality is due to Jensen's inequality. Now we will bound B , using the safe separability measurement σ' as defined in (11).

$$\begin{aligned} B &= \left\| \frac{1}{\lambda n} A (\alpha + \gamma \sum_{k=1}^K \Delta \alpha_{[k]}) \right\|^2 = \left\| \mathbf{w}(\alpha) + \gamma \frac{1}{\lambda n} \sum_{k=1}^K A \Delta \alpha_{[k]} \right\|^2 \\ &= \|\mathbf{w}(\alpha)\|^2 + \sum_{k=1}^K 2\gamma \frac{1}{\lambda n} \mathbf{w}(\alpha)^T A \Delta \alpha_{[k]} + \gamma \left(\frac{1}{\lambda n} \right)^2 \gamma \left\| \sum_{k=1}^K A \Delta \alpha_{[k]} \right\|^2 \\ &\stackrel{(11)}{\leq} \|\mathbf{w}(\alpha)\|^2 + \sum_{k=1}^K 2\gamma \frac{1}{\lambda n} \mathbf{w}(\alpha)^T A \Delta \alpha_{[k]} + \gamma \left(\frac{1}{\lambda n} \right)^2 \sigma' \sum_{k=1}^K \|A \Delta \alpha_{[k]}\|^2. \end{aligned}$$

Plugging A and B into (25) will give us

$$\begin{aligned}
 \mathcal{D}(\boldsymbol{\alpha} + \gamma \sum_{k=1}^K \Delta \boldsymbol{\alpha}_{[k]}) &\geq -\frac{1}{n} \sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} (1-\gamma) \ell_i^*(-\alpha_i) + \gamma \ell_i^*(-(\boldsymbol{\alpha} + \Delta \boldsymbol{\alpha}_{[k]})_i) \right) \\
 &\quad - \gamma \frac{\lambda}{2} \|\mathbf{w}(\boldsymbol{\alpha})\|^2 - (1-\gamma) \frac{\lambda}{2} \|\mathbf{w}(\boldsymbol{\alpha})\|^2 - \frac{\lambda}{2} \sum_{k=1}^K 2\gamma \frac{1}{\lambda n} \mathbf{w}(\boldsymbol{\alpha})^T A \Delta \boldsymbol{\alpha}_{[k]} - \frac{\lambda}{2} \gamma \left(\frac{1}{\lambda n} \right)^2 \sigma' \sum_{k=1}^K \|A \Delta \boldsymbol{\alpha}_{[k]}\|^2 \\
 &= -\underbrace{\frac{1}{n} \sum_{k=1}^K \left(\sum_{i \in \mathcal{P}_k} (1-\gamma) \ell_i^*(-\alpha_i) \right)}_{(1-\gamma)\mathcal{D}(\boldsymbol{\alpha})} - (1-\gamma) \frac{\lambda}{2} \|\mathbf{w}(\boldsymbol{\alpha})\|^2 \\
 &\quad + \gamma \sum_{k=1}^K \left(-\frac{1}{n} \sum_{i \in \mathcal{P}_k} \ell_i^*(-(\boldsymbol{\alpha} + \Delta \boldsymbol{\alpha}_{[k]})_i) - \frac{1}{K} \frac{\lambda}{2} \|\mathbf{w}(\boldsymbol{\alpha})\|^2 - \frac{1}{n} \mathbf{w}(\boldsymbol{\alpha})^T A \Delta \boldsymbol{\alpha}_{[k]} - \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A \Delta \boldsymbol{\alpha}_{[k]} \right\|^2 \right) \\
 &\stackrel{(9)}{=} (1-\gamma)\mathcal{D}(\boldsymbol{\alpha}) + \gamma \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}; \mathbf{w}, \boldsymbol{\alpha}_{[k]}).
 \end{aligned}$$

B.2. Proof of Lemma 4

See (Richtárik & Takáč, 2013).

B.3. Proof of Lemma 5

For sake of notation, we will write $\boldsymbol{\alpha}$ instead of $\boldsymbol{\alpha}^{(t)}$, \mathbf{w} instead of $\mathbf{w}(\boldsymbol{\alpha}^{(t)})$ and \mathbf{u} instead of $\mathbf{u}^{(t)}$.

Now, let us estimate the expected change of the dual objective. Using the definition of the dual update $\boldsymbol{\alpha}^{(t+1)} := \boldsymbol{\alpha}^{(t)} + \gamma \sum_k \Delta \boldsymbol{\alpha}_{[k]}$ resulting in Algorithm 1, we have

$$\begin{aligned}
 \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)})] &= \mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha} + \gamma \sum_{k=1}^K \Delta \boldsymbol{\alpha}_{[k]})\right] \\
 &\quad \text{(by Lemma 3 on the local function } \mathcal{G}_k^{\sigma'}(\boldsymbol{\alpha}; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) \text{ approximating the global objective } \mathcal{D}(\boldsymbol{\alpha})) \\
 &\leq \mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}) - (1-\gamma)\mathcal{D}(\boldsymbol{\alpha}) - \gamma \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^{(t)}; \mathbf{w}, \boldsymbol{\alpha}_{[k]})\right] \\
 &= \gamma \mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^{(t)}; \mathbf{w}, \boldsymbol{\alpha}_{[k]})\right] \\
 &= \gamma \mathbb{E}\left[\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^*; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) + \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^*; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) - \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^{(t)}; \mathbf{w}, \boldsymbol{\alpha}_{[k]})\right] \\
 &\quad \text{(by the notion of quality (12) of the local solver, as in Assumption 1)} \\
 &\leq \gamma \left(\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^*; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) + \underbrace{\Theta \left(\sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^*; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) - \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\mathbf{0}; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) \right)}_{\mathcal{D}(\boldsymbol{\alpha})} \right) \\
 &= \gamma (1-\Theta) \underbrace{\left(\mathcal{D}(\boldsymbol{\alpha}) - \sum_{k=1}^K \mathcal{G}_k^{\sigma'}(\Delta \boldsymbol{\alpha}_{[k]}^*; \mathbf{w}, \boldsymbol{\alpha}_{[k]}) \right)}_C. \tag{26}
 \end{aligned}$$

Now, let us upper bound the C term (we will denote by $\Delta\alpha^* = \sum_{k=1}^K \Delta\alpha_{[k]}^*$):

$$\begin{aligned}
 C &\stackrel{(2),(9)}{=} \frac{1}{n} \sum_{i=1}^n (\ell_i^*(-\alpha_i - \Delta\alpha_i^*) - \ell_i^*(-\alpha_i)) + \frac{1}{n} \mathbf{w}(\alpha)^T A \Delta\alpha^* + \sum_{k=1}^K \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A \Delta\alpha_{[k]}^* \right\|^2 \\
 &\leq \frac{1}{n} \sum_{i=1}^n (\ell_i^*(-\alpha_i - s(u_i - \alpha_i)) - \ell_i^*(-\alpha_i)) + \frac{1}{n} \mathbf{w}(\alpha)^T A s(\mathbf{u} - \alpha) + \sum_{k=1}^K \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A s(\mathbf{u} - \alpha)_{[k]} \right\|^2 \\
 &\stackrel{\text{Strong conv.}}{\leq} \frac{1}{n} \sum_{i=1}^n \left(s\ell_i^*(-u_i) + (1-s)\ell_i^*(-\alpha_i) - \frac{\mu}{2}(1-s)s(u_i - \alpha_i)^2 - \ell_i^*(-\alpha_i) \right) + \frac{1}{n} \mathbf{w}(\alpha)^T A s(\mathbf{u} - \alpha) \\
 &\quad + \sum_{k=1}^K \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A s(\mathbf{u} - \alpha)_{[k]} \right\|^2 \\
 &= \frac{1}{n} \sum_{i=1}^n \left(s\ell_i^*(-u_i) - s\ell_i^*(-\alpha_i) - \frac{\mu}{2}(1-s)s(u_i - \alpha_i)^2 \right) + \frac{1}{n} \mathbf{w}(\alpha)^T A s(\mathbf{u} - \alpha) + \sum_{k=1}^K \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A s(\mathbf{u} - \alpha)_{[k]} \right\|^2.
 \end{aligned}$$

The convex conjugate maximal property implies that

$$\ell_i^*(-u_i) = -u_i \mathbf{w}(\alpha)^T \mathbf{x}_i - \ell_i(\mathbf{w}(\alpha)^T \mathbf{x}_i). \quad (27)$$

Moreover, from the definition of the primal and dual optimization problems (1), (2), we can write the duality gap as

$$G(\alpha) := \mathcal{P}(\mathbf{w}(\alpha)) - \mathcal{D}(\alpha) \stackrel{(1),(2)}{=} \frac{1}{n} \sum_{i=1}^n (\ell_i(\mathbf{x}_i^T \mathbf{w}) + \ell_i^*(-\alpha_i) + \mathbf{w}(\alpha)^T \mathbf{x}_i \alpha_i). \quad (28)$$

Hence,

$$\begin{aligned}
 C &\stackrel{(27)}{\leq} \frac{1}{n} \sum_{i=1}^n \left(-su_i \mathbf{w}(\alpha)^T \mathbf{x}_i - s\ell_i(\mathbf{w}(\alpha)^T \mathbf{x}_i) - s\ell_i^*(-\alpha_i) - \underbrace{-s\mathbf{w}(\alpha)^T \mathbf{x}_i \alpha_i + s\mathbf{w}(\alpha)^T \mathbf{x}_i \alpha_i}_0 - \frac{\mu}{2}(1-s)s(u_i - \alpha_i)^2 \right) \\
 &\quad + \frac{1}{n} \mathbf{w}(\alpha)^T A s(\mathbf{u} - \alpha) + \sum_{k=1}^K \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A s(\mathbf{u} - \alpha)_{[k]} \right\|^2 \\
 &= \frac{1}{n} \sum_{i=1}^n (-s\ell_i(\mathbf{w}(\alpha)^T \mathbf{x}_i) - s\ell_i^*(-\alpha_i) - s\mathbf{w}(\alpha)^T \mathbf{x}_i \alpha_i) + \frac{1}{n} \sum_{i=1}^n \left(s\mathbf{w}(\alpha)^T \mathbf{x}_i (\alpha_i - u_i) - \frac{\mu}{2}(1-s)s(u_i - \alpha_i)^2 \right) \\
 &\quad + \frac{1}{n} \mathbf{w}(\alpha)^T A s(\mathbf{u} - \alpha) + \sum_{k=1}^K \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A s(\mathbf{u} - \alpha)_{[k]} \right\|^2 \\
 &\stackrel{(28)}{=} -sG(\alpha) - \frac{\mu}{2}(1-s)s \frac{1}{n} \sum_{i=1}^n \|\mathbf{u} - \alpha\|^2 + \frac{\sigma'}{2\lambda} \left(\frac{s}{n} \right)^2 \sum_{k=1}^K \|A(\mathbf{u} - \alpha)_{[k]}\|^2. \quad (29)
 \end{aligned}$$

Now, the claimed improvement bound (15) follows by plugging (29) into (26).

B.4. Proof of Lemma 6

For general convex functions, the strong convexity parameter is $\mu = 0$, and hence the definition of $R^{(t)}$ becomes

$$R^{(t)} \stackrel{(16)}{=} \sum_{k=1}^K \|A(\mathbf{u}^{(t)} - \alpha^{(t)})_{[k]}\|^2 \stackrel{(19)}{\leq} \sum_{k=1}^K \sigma_k \|(\mathbf{u}^{(t)} - \alpha^{(t)})_{[k]}\|^2 \stackrel{\text{Lemma 16}}{\leq} \sum_{k=1}^K \sigma_k |\mathcal{P}_k| 4L^2.$$

B.5. Proof of Theorem 8

At first let us estimate expected change of dual feasibility. By using the main Lemma 5, we have

$$\begin{aligned}
 \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)})] &= \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)}) + \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \\
 &\stackrel{(15)}{=} \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \gamma(1 - \Theta)sG(\boldsymbol{\alpha}^{(t)}) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 R^{(t)} \\
 &\stackrel{(4)}{=} \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \gamma(1 - \Theta)s(\mathcal{P}(\mathbf{w}(\boldsymbol{\alpha}^{(t)})) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 R^{(t)} \\
 &\leq \mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)}) - \gamma(1 - \Theta)s(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 R^{(t)} \\
 &\stackrel{(18)}{\leq} (1 - \gamma(1 - \Theta)s)(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 4L^2\sigma.
 \end{aligned} \tag{30}$$

Using (30) recursively we have

$$\begin{aligned}
 \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] &= (1 - \gamma(1 - \Theta)s)^t (\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 4L^2\sigma \sum_{j=0}^{t-1} (1 - \gamma(1 - \Theta)s)^j \\
 &= (1 - \gamma(1 - \Theta)s)^t (\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 4L^2\sigma \frac{1 - (1 - \gamma(1 - \Theta)s)^t}{\gamma(1 - \Theta)s} \\
 &\leq (1 - \gamma(1 - \Theta)s)^t (\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})) + s \frac{4L^2\sigma\sigma'}{2\lambda n^2}.
 \end{aligned} \tag{31}$$

Choice of $s = 1$ and $t = t_0 := \max\{0, \lceil \frac{1}{\gamma(1-\Theta)} \log(2\lambda n^2(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})) / (4L^2\sigma\sigma')) \rceil\}$ will lead to

$$\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \leq (1 - \gamma(1 - \Theta))^{t_0} (\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(0)})) + \frac{4L^2\sigma\sigma'}{2\lambda n^2} \leq \frac{4L^2\sigma\sigma'}{2\lambda n^2} + \frac{4L^2\sigma\sigma'}{2\lambda n^2} = \frac{4L^2\sigma\sigma'}{\lambda n^2}. \tag{32}$$

Now, we are going to show that

$$\forall t \geq t_0 : \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})] \leq \frac{4L^2\sigma\sigma'}{\lambda n^2(1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0))}. \tag{33}$$

Clearly, (32) implies that (33) holds for $t = t_0$. Now imagine that it holds for any $t \geq t_0$ then we show that it also has to hold for $t + 1$. Indeed, using

$$s = \frac{1}{1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0)} \in [0, 1] \tag{34}$$

we obtain

$$\begin{aligned}
 \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t+1)})] &\stackrel{(30)}{\leq} (1 - \gamma(1 - \Theta)s)(\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}^{(t)})) + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 4L^2\sigma \\
 &\stackrel{(33)}{\leq} (1 - \gamma(1 - \Theta)s) \frac{4L^2\sigma\sigma'}{\lambda n^2(1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0))} + \gamma(1 - \Theta)\frac{\sigma'}{2\lambda}\left(\frac{s}{n}\right)^2 4L^2\sigma \\
 &\stackrel{(34)}{=} \frac{4L^2\sigma\sigma'}{\lambda n^2} \left(\frac{1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0) - \gamma(1 - \Theta) + \gamma(1 - \Theta)\frac{1}{2}}{(1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0))^2} \right) \\
 &= \frac{4L^2\sigma\sigma'}{\lambda n^2} \underbrace{\left(\frac{1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0) - \frac{1}{2}\gamma(1 - \Theta)}{(1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0))^2} \right)}_D.
 \end{aligned}$$

Now, we will upperbound D as follows

$$\begin{aligned}
 D &= \frac{1}{1 + \frac{1}{2}\gamma(1 - \Theta)(t + 1 - t_0)} \underbrace{\frac{(1 + \frac{1}{2}\gamma(1 - \Theta)(t + 1 - t_0))(1 + \frac{1}{2}\gamma(1 - \Theta)(t - 1 - t_0))}{(1 + \frac{1}{2}\gamma(1 - \Theta)(t - t_0))^2}}_{\leq 1} \\
 &\leq \frac{1}{1 + \frac{1}{2}\gamma(1 - \Theta)(t + 1 - t_0)},
 \end{aligned}$$

where in the last inequality we have used the fact that geometric mean is less or equal to arithmetic mean.

If $\bar{\alpha}$ is defined as (21) then we obtain that

$$\begin{aligned}
 \mathbb{E}[G(\bar{\alpha})] &= \mathbb{E} \left[G \left(\sum_{t=T_0}^{T-1} \frac{1}{T-T_0} \alpha^{(t)} \right) \right] \leq \frac{1}{T-T_0} \mathbb{E} \left[\sum_{t=T_0}^{T-1} G(\alpha^{(t)}) \right] \\
 &\stackrel{(15),(18)}{\leq} \frac{1}{T-T_0} \mathbb{E} \left[\sum_{t=T_0}^{T-1} \left(\frac{1}{\gamma(1-\Theta)s} (\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^{(t)})) + \frac{4L^2\sigma\sigma's}{2\lambda n^2} \right) \right] \\
 &= \frac{1}{\gamma(1-\Theta)s} \frac{1}{T-T_0} \mathbb{E} [\mathcal{D}(\alpha^{(T)}) - \mathcal{D}(\alpha^{(T_0)})] + \frac{4L^2\sigma\sigma's}{2\lambda n^2} \\
 &\leq \frac{1}{\gamma(1-\Theta)s} \frac{1}{T-T_0} \mathbb{E} [\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(T_0)})] + \frac{4L^2\sigma\sigma's}{2\lambda n^2}.
 \end{aligned} \tag{35}$$

Now, if $T \geq \lceil \frac{1}{\gamma(1-\Theta)} \rceil + T_0$ such that $T_0 \geq t_0$ we obtain

$$\begin{aligned}
 \mathbb{E}[G(\bar{\alpha})] &\stackrel{(35),(33)}{\leq} \frac{1}{\gamma(1-\Theta)s} \frac{1}{T-T_0} \left(\frac{4L^2\sigma\sigma'}{\lambda n^2(1 + \frac{1}{2}\gamma(1-\Theta)(T_0 - t_0))} \right) + \frac{4L^2\sigma\sigma's}{2\lambda n^2} \\
 &= \frac{4L^2\sigma\sigma'}{\lambda n^2} \left(\frac{1}{\gamma(1-\Theta)s} \frac{1}{T-T_0} \frac{1}{1 + \frac{1}{2}\gamma(1-\Theta)(T_0 - t_0)} + \frac{s}{2} \right).
 \end{aligned} \tag{36}$$

Choosing

$$s = \frac{1}{(T - T_0)\gamma(1 - \Theta)} \in [0, 1] \tag{37}$$

gives us

$$\mathbb{E}[G(\bar{\alpha})] \stackrel{(36),(37)}{\leq} \frac{4L^2\sigma\sigma'}{\lambda n^2} \left(\frac{1}{1 + \frac{1}{2}\gamma(1-\Theta)(T_0 - t_0)} + \frac{1}{(T - T_0)\gamma(1 - \Theta)} \frac{1}{2} \right). \tag{38}$$

To have right hand side of (38) smaller then ϵ_G it is sufficient to choose T_0 and T such that

$$\frac{4L^2\sigma\sigma'}{\lambda n^2} \left(\frac{1}{1 + \frac{1}{2}\gamma(1-\Theta)(T_0 - t_0)} \right) \leq \frac{1}{2}\epsilon_G, \tag{39}$$

$$\frac{4L^2\sigma\sigma'}{\lambda n^2} \left(\frac{1}{(T - T_0)\gamma(1 - \Theta)} \frac{1}{2} \right) \leq \frac{1}{2}\epsilon_G. \tag{40}$$

Hence of if

$$\begin{aligned}
 t_0 + \frac{2}{\gamma(1-\Theta)} \left(\frac{8L^2\sigma\sigma'}{\lambda n^2\epsilon_G} - 1 \right) &\leq T_0, \\
 T_0 + \frac{4L^2\sigma\sigma'}{\lambda n^2\epsilon_G\gamma(1-\Theta)} &\leq T,
 \end{aligned}$$

then (39) and (40) are satisfied.

B.6. Proof of Theorem 10

If the function $\ell_i(\cdot)$ is $(1/\mu)$ -smooth then $\ell_i^*(\cdot)$ is μ -strongly convex with respect to the $\|\cdot\|$ norm. From (16) we have

$$\begin{aligned}
 R^{(t)} &\stackrel{(16)}{=} -\frac{\lambda\mu n(1-s)}{\sigma's} \|\mathbf{u}^{(t)} - \alpha^{(t)}\|^2 + \sum_{k=1}^K \|A(\mathbf{u}^{(t)} - \alpha^{(t)})_{[k]}\|^2 \\
 &\stackrel{(19)}{\leq} -\frac{\lambda\mu n(1-s)}{\sigma's} \|\mathbf{u}^{(t)} - \alpha^{(t)}\|^2 + \sum_{k=1}^K \sigma_k \|\mathbf{u}^{(t)} - \alpha_{[k]}^{(t)}\|^2 \\
 &\leq -\frac{\lambda\mu n(1-s)}{\sigma's} \|\mathbf{u}^{(t)} - \alpha^{(t)}\|^2 + \sigma_{\max} \sum_{k=1}^K \|\mathbf{u}^{(t)} - \alpha_{[k]}^{(t)}\|^2 \\
 &= \left(-\frac{\lambda\mu n(1-s)}{\sigma's} + \sigma_{\max} \right) \|\mathbf{u}^{(t)} - \alpha^{(t)}\|^2.
 \end{aligned} \tag{41}$$

If we plug

$$s = \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'} \in [0, 1] \quad (42)$$

into (41) we obtain that $\forall t : R^{(t)} \leq 0$. Putting the same s into (15) will give us

$$\mathbb{E}[\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^{(t)})] \stackrel{(15),(42)}{\geq} \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'} G(\alpha^{(t)}) \geq \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'} \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)}). \quad (43)$$

Using the fact that $\mathbb{E}[\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^{(t)})] = \mathbb{E}[\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^*)] + \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)})$ we have

$$\mathbb{E}[\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^*)] + \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)}) \stackrel{(43)}{\geq} \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'} \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)})$$

which is equivalent with

$$\mathbb{E}[\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t+1)})] \leq \left(1 - \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'}\right) \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)}). \quad (44)$$

Therefore if we denote by $\epsilon_{\mathcal{D}}^{(t)} = \mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)})$ we have that

$$\mathbb{E}[\epsilon_{\mathcal{D}}^{(t)}] \stackrel{(44)}{\leq} \left(1 - \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'}\right)^t \epsilon_{\mathcal{D}}^{(0)} \stackrel{(24)}{\leq} \left(1 - \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'}\right)^t \leq \exp\left(-t\gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'}\right).$$

The right hand side will be smaller than some $\epsilon_{\mathcal{D}}$ if

$$t \geq \frac{1}{\gamma(1 - \Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \log \frac{1}{\epsilon_{\mathcal{D}}}.$$

Moreover, to bound the duality gap, we have

$$\gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'} G(\alpha^{(t)}) \stackrel{(43)}{\leq} \mathbb{E}[\mathcal{D}(\alpha^{(t+1)}) - \mathcal{D}(\alpha^{(t)})] \leq \mathbb{E}[\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha^{(t)})].$$

Therefore $G(\alpha^{(t)}) \leq \frac{1}{\gamma(1 - \Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \epsilon_{\mathcal{D}}^{(t)}$. Hence if $\epsilon_{\mathcal{D}} \leq \gamma(1 - \Theta) \frac{\lambda\mu n}{\lambda\mu n + \sigma_{\max}\sigma'} \epsilon_G$ then $G(\alpha^{(t)}) \leq \epsilon_G$. Therefore after

$$t \geq \frac{1}{\gamma(1 - \Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \log \left(\frac{1}{\gamma(1 - \Theta)} \frac{\lambda\mu n + \sigma_{\max}\sigma'}{\lambda\mu n} \frac{1}{\epsilon_G} \right)$$

iterations we have obtained a duality gap less than ϵ_G .

B.7. Proof of Theorem 13

Because ℓ_i are $(1/\mu)$ -smooth then functions ℓ_i^* are μ strongly convex with respect to the norm $\|\cdot\|$. The proof is based on techniques developed in recent coordinate descent papers, including (Richtárik & Takáč, 2014; 2013; Richtárik & Takáč, 2015; Tappenden et al., 2015; Mareček et al., 2014; Fercoq & Richtárik, 2013; Lu & Xiao, 2013; Fercoq et al., 2014; Qu & Richtárik, 2014; Qu et al., 2014) (Efficient accelerated variants were considered in (Fercoq & Richtárik, 2013; Shalev-Shwartz & Zhang, 2013a)).

First, let us define the function $F(\zeta) : \mathbb{R}^{n_k} \rightarrow \mathbb{R}$ as $F(\zeta) := -\mathcal{G}_k^{\sigma'}(\sum_{i \in \mathcal{P}_k} \zeta_i \mathbf{e}_i; \mathbf{w}, \alpha_{[k]})$. This function can be written in two parts $F(\zeta) = \Phi(\zeta) + f(\zeta)$. The first part denoted by $\Phi(\zeta) = \frac{1}{n} \sum_{i \in \mathcal{P}_k} \ell_i^*(-\alpha_i - \zeta_i)$ is strongly convex with convexity parameter $\frac{\mu}{n}$ with respect to the standard Euclidean norm. In our application, we think of the ζ variable collecting the local dual variables $\Delta\alpha_{[k]}$.

The second part we will denote by $f(\zeta) = \frac{1}{K} \frac{\lambda}{2} \|\mathbf{w}(\alpha)\|^2 + \frac{1}{n} \sum_{i \in \mathcal{P}_k} \mathbf{w}(\alpha)^T \mathbf{x}_i \zeta_i + \frac{\lambda}{2} \sigma' \frac{1}{\lambda^2 n^2} \|\sum_{i \in \mathcal{P}_k} \mathbf{x}_i \zeta_i\|^2$. It is easy to show that the gradient of f is coordinate-wise Lipschitz continuous with Lipschitz constant $\frac{\sigma'}{\lambda n^2} r_{\max}$ with respect to the standard Euclidean norm.

Following the proof of Theorem 20 in (Richtárik & Takáč, 2015), we obtain that

$$\begin{aligned} \mathbb{E}[\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^{(h+1)}; \mathbf{w}, \alpha_{[k]})] &\leq \left(1 - \frac{1}{n_k} \frac{1 + \frac{\mu n \lambda}{\sigma' r_{\max}}}{\frac{\mu n \lambda}{\sigma' r_{\max}}}\right) \left(\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^{(h)}; \mathbf{w}, \alpha_{[k]})\right) \\ &= \left(1 - \frac{1}{n_k} \frac{\lambda n \mu}{\sigma' r_{\max} + \lambda n \mu}\right) \left(\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^{(h)}; \mathbf{w}, \alpha_{[k]})\right). \end{aligned}$$

Over all steps up to step h , this gives

$$\mathbb{E}[\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^{(h)}; \mathbf{w}, \alpha_{[k]})] \leq \left(1 - \frac{1}{n_k} \frac{\lambda n \mu}{\sigma' r_{\max} + \lambda n \mu}\right)^h \left(\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\mathbf{0}; \mathbf{w}, \alpha_{[k]})\right).$$

Therefore, choosing H as in the assumption of our Theorem, given in Equation (22), we are guaranteed that $\left(1 - \frac{1}{n_k} \frac{\lambda n \mu}{\sigma' r_{\max} + \lambda n \mu}\right)^H \leq \Theta$, as desired.

B.8. Proof of Theorem 14

Similarly as in the proof of Theorem 13 we define a composite function $F(\zeta) = f(\zeta) + \Phi(\zeta)$. However, in this case functions ℓ_i^* are not guaranteed to be strongly convex. However, the first part has still a coordinate-wise Lipschitz continuous gradient with constant $\frac{\sigma'}{\lambda n^2} r_{\max}$ with respect to the standard Euclidean norm. Therefore from Theorem 3 in (Tappenden et al., 2015) we have that

$$\mathbb{E}[\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^{(h)}; \mathbf{w}, \alpha_{[k]})] \leq \frac{n_k}{n_k + h} \left(\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\mathbf{0}; \mathbf{w}, \alpha_{[k]}) + \frac{1}{2} \frac{\sigma' r_{\max}}{\lambda n^2} \|\Delta\alpha_{[k]}^*\|^2\right). \quad (45)$$

Now, choice of $h = H$ from (23) is sufficient to have the right hand side of (45) to be $\leq \Theta(\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}^*; \mathbf{w}, \alpha_{[k]}) - \mathcal{G}_k^{\sigma'}(\mathbf{0}; \mathbf{w}, \alpha_{[k]}))$.

C. Relationship of DisDCA to CoCoA⁺

We are indebted to Ching-pei Lee for showing the following relationship between the practical variant of DisDCA (Yang, 2013), and CoCoA⁺ when SDCA is chosen as the local solver:

Considering the practical variant of DisDCA (DisDCA-p, see Figure 2 in (Yang, 2013)) using the scaling parameter $scl = K$, the following holds:

Lemma 18. *Assume that the dataset is partitioned equally between workers, i.e. $\forall k : n_k = \frac{n}{K}$. If within the CoCoA⁺ framework, SDCA is used as a local solver, and the subproblems are formulated using our shown “safe” (but pessimistic) upper bound of $\sigma' = K$, with aggregation parameter $\gamma = 1$ (adding), then the CoCoA⁺ framework reduces exactly to the DisDCA-p algorithm.*

Proof. (Due to Ching-pei Lee, with some reformulations). As defined in (9), the data-local subproblem solved by each machine in CoCoA⁺ is defined as

$$\max_{\Delta\alpha_{[k]} \in \mathbb{R}^n} \mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}; \mathbf{w}, \alpha_{[k]})$$

where

$$\mathcal{G}_k^{\sigma'}(\Delta\alpha_{[k]}; \mathbf{w}, \alpha_{[k]}) := -\frac{1}{n} \sum_{i \in \mathcal{P}_k} \ell_i^*(-\alpha_i - (\Delta\alpha_{[k]})_i) - \frac{1}{K} \frac{\lambda}{2} \|\mathbf{w}\|^2 - \frac{1}{n} \mathbf{w}^T A \Delta\alpha_{[k]} - \frac{\lambda}{2} \sigma' \left\| \frac{1}{\lambda n} A \Delta\alpha_{[k]} \right\|^2.$$

We rewrite the local problem by scaling with n , and removing the constant regularizer term $\frac{1}{K} \frac{\lambda}{2} \|\mathbf{w}\|^2$, i.e.

$$\tilde{\mathcal{G}}_k^{\sigma'}(\Delta\alpha_{[k]}; \mathbf{w}) := - \sum_{j \in \mathcal{P}_k} \ell_j^*(-\alpha_j - (\Delta\alpha_{[k]})_j) - \mathbf{w}^T A \Delta\alpha_{[k]} - \frac{\sigma'}{2\lambda n} \left\| A \Delta\alpha_{[k]} \right\|^2. \quad (46)$$

For the correspondence of interest, we now restrict to single coordinate updates in the local solver. In other words, the local solver optimizes exactly one coordinate $i \in \mathcal{P}_k$ at a time. To relate the single coordinate update to the set of local variables, we will use the notation

$$\Delta\alpha_{[k]} =: \Delta\alpha_{[k]}^{\text{prev}} + \delta\mathbf{e}_i, \quad (47)$$

so that $\Delta\alpha_{[k]}^{\text{prev}}$ are the previous local variables, and $\Delta\alpha_{[k]}$ will be the updated ones.

From now on, we will consider the special case of CoCoA^+ when the quadratic upper bound parameter is chosen as the “safe” value $\sigma' = K$, combined with adding as the aggregation, i.e. $\gamma = 1$.

Now if the local solver within CoCoA^+ is chosen as LOCALSDCA , then one local step on the subproblem (9) will calculate the following coordinate update. Recall that $A = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$.

$$\delta^* := \arg \max_{\delta \in \mathbb{R}} \tilde{\mathcal{G}}_k^{\sigma'}(\Delta\alpha_{[k]}; \mathbf{w}) \quad (48)$$

which – because it is only affecting one single coordinate, employing (47) – can be expressed as

$$\begin{aligned} \delta^* &:= \arg \max_{\delta \in \mathbb{R}} -\ell_i^*(-(\alpha_i + (\Delta\alpha_{[k]}^{\text{prev}})_i + \delta)) - \delta \mathbf{x}_i^T \mathbf{w} - \frac{K}{\lambda n} \delta \mathbf{x}_i^T A \Delta\alpha_{[k]}^{\text{prev}} - \frac{K}{2\lambda n} \delta^2 \|\mathbf{x}_i\|_2^2 \\ &= \arg \max_{\delta \in \mathbb{R}} -\ell_i^*(-(\alpha_i + (\Delta\alpha_{[k]}^{\text{prev}})_i + \delta)) - \delta \mathbf{x}_i^T \underbrace{\left(\mathbf{w} + \frac{K}{\lambda n} A \Delta\alpha_{[k]}^{\text{prev}} \right)}_{=: \mathbf{u}^{\text{local}}} - \frac{K}{2\lambda n} \delta^2 \|\mathbf{x}_i\|_2^2 \end{aligned} \quad (49)$$

From this formulation, it is apparent that single coordinate local solvers should maintain their locally updated version of the current primal parameters, which we here denote as

$$\mathbf{u}^{\text{local}} = \mathbf{w} + \frac{K}{\lambda n} A \Delta\alpha_{[k]}^{\text{prev}}. \quad (50)$$

In the practical variant of DisDCA , the summarized local primal updates are $\Delta\mathbf{u}^{\text{local}} = \frac{1}{\lambda n_k} A \Delta\alpha_{[k]}$. For the balanced case $n_k = n/K$ for K being the number of machines, this means the local $\mathbf{u}^{\text{local}}$ update of DisDCA-p is

$$\Delta\alpha_i^* := \arg \max_{\Delta\alpha_i \in \mathbb{R}} -\ell_i^*(-(\alpha_i + \Delta\alpha_i)) - \Delta\alpha_i \mathbf{x}_i^T \mathbf{u}^{\text{local}} - \frac{K}{2\lambda n} (\Delta\alpha_i)^2 \|\mathbf{x}_i\|_2^2. \quad (51)$$

It is not hard to show that during one outer round, the evolution of the local dual variables $\Delta\alpha_{[k]}$ is the same in both methods, such that they will also have the same trajectory of $\mathbf{u}^{\text{local}}$. This requires some care if the same coordinate is sampled more than once in a round, which can happen in LOCALSDCA within CoCoA^+ and also in DisDCA-p . \square

Discussion. In the view of the above lemma, we will summarize the connection of the two methods as follows:

- **CoCoA/+ is Not an Algorithm.** In contrast, it is a framework which allows to use *any local solver* to perform approximate steps on the local subproblem. This additional level of abstraction (from the definition of such local subproblems in (9)) is the first to allow *reusability* of any fast/tuned and problem specific single machine solvers, while decoupling this from the distributed algorithmic scheme, as presented in Algorithm 1.

Concerning the choice of local solver to be used within CoCoA/+ , SDCA is *not* the fastest known single machine solver for most applications. Much recent research has shown improvements on SDCA (Shalev-Shwartz & Zhang, 2013c), such as accelerated variants (Shalev-Shwartz & Zhang, 2013b) and other approaches including variance reduction, methods incorporating second-order information, and importance sampling. In this light, we encourage the user of the CoCoA or CoCoA^+ framework to plug in the best and most recent solver available for their particular local problem (within Algorithm 1), which is not necessarily SDCA . This choice should be made explicit especially when comparing algorithms. Our presented convergence theory from Section 4 will still cover these choices, since it only depends on the relative accuracy Θ of the chosen local solver.

- **CoCoA⁺ is Theoretically Safe, while still Adaptive to the Data.** The general definition of the local subproblems, and therefore the treatment of the varying separable bound on the objective – quantified by σ' – allows our framework to adapt to the difficulty of the data partition and still give convergence results. The data-dependent measure σ' is fully decoupled from what the user of the framework prefers to employ as a local solver (see also the comment below that CoCoA is not a coordinate solver).

The safe upper bound $\sigma' = K$ is worst-case pessimistic, for the convergence theory to still hold in all cases, when the updates are added. Using additional knowledge from the input data, better bounds and therefore better step-sizes can be achieved in CoCoA⁺. An example when σ' can be safely chosen much smaller is when the data-matrix satisfies strong row/column sparsity, see e.g. Lemma 1 in (Richtárik & Takáč, 2013).

- **Obtaining DisDCA-p as a Special Case.** As shown in Lemma 18 above, we have that if in CoCoA⁺, if SDCA is used as the local solver and the pessimistic upper bound of $\sigma' = K$ is used and, moreover, the dataset is partitioned equally, i.e. $\forall k : n_k = \frac{n}{K}$, then the CoCoA⁺ framework reduces exactly to the DisDCA-p algorithm by (Yang, 2013).

The correspondence breaks down if the subproblem parameter is chosen to a practically good value $\sigma' \neq K$. Also, as noted above, SDCA is often not the best local solver currently available. In our above experiments, SDCA was used just for demonstration purposes and ease of comparison. Furthermore, the data partition might often be unbalanced in practical applications.

While both DisDCA-p and CoCoA are special cases of CoCoA⁺, we note that DisDCA-p can not be recovered as a special case of the original CoCoA framework (Jaggi et al., 2014).

- **CoCoA/+ are Not Coordinate Methods.** Despite the original name being motivated from this special case, CoCoA and CoCoA⁺ are *not* coordinate methods. In fact, CoCoA⁺ as presented here for the adding case ($\gamma = 1$) is much more closely related to a batch method applied to the dual, using a block-separable proximal term, as following from our new subproblem formulation (9), depending on σ' . See also the remark in Section 6. The framework here (Algorithm 1) gives more generality, as the used local solver is not restricted to be a coordinate-wise one. In fact the framework allows to translate recent and future improvements of single machine solvers directly to the distributed setting, by employing them within Algorithm 1. DisDCA-p works very well for several applications, but is restricted to using local coordinate ascent (SDCA) steps.

- **Theoretical Convergence Results.** While DisDCA-p (Yang, 2013) was proposed without theoretical justification (hence the nomenclature), the main contribution in the paper here – apart from the arbitrary local solvers – is the convergence analysis for the framework. The theory proposed in (Yang et al., 2013) is given only for the setting of orthogonal partitions, i.e., when $\sigma' = 1$ and the problems become trivial to distribute given the orthogonality of data between the workers.

The theoretical analysis here gives convergence rates applying for Algorithm 1 when using arbitrary local solvers, and inherits the performance of the local solver. As a special case, we obtain the first theoretical justification and convergence rates for original CoCoA in the case of general convex objective, as well as for the special case of DisDCA-p for both general convex and smooth convex objectives.